(b) generating permuted responses of genes by means of Monte Carlo randomization of perturbation index for the response of each gene across all perturbations;

(c) performing cluster analysis on the permuted responses of genes;

(d) determining for each cluster generated in step (c) the fractional improvement in the cluster analysis of genes based on the permuted responses of genes, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters; and

(e) repeating steps (b) through (d) so that a distribution of fractional improvements in the cluster analysis of the genes is obtained for each cluster generated by said cluster analysis;

wherein the statistical significance of each of said sets of co-varying genes is determined by comparing the actual fractional improvement for the corresponding cluster to the distribution of fractional improvements for the corresponding cluster.

REMARKS

Claims 1, 3-50, 58-64, 72-78, 89-100 and 105-124 are pending in the application. In the instant Amendment, claims 1, 18, 26, 29, 38, 50, 64, 96, 106 and 123 have been amended to clarify the present invention. Upon entry of the above-made amendments, claims 1, 3-50, 58-64, 72-78, 89-100 and 105-124 will be pending. A marked version showing changes made to the amended claims is attached hereto as Exhibit A. A clean version of the pending claims, as amended, is attached hereto as Exhibit B.

Claim 1 has been amended to recite that the claimed methods comprise *determining, for each of a plurality of sets of cellular constituents in a plurality of response profiles, whether said set of cellular constituents is upregulated or downregulated by said first plurality of drug perturbations*, and that the consensus profile for said first plurality of drug perturbations comprises *measurements of said set or sets of cellular constituents that are determined in said determining step to be upregulated or downregulated by said first plurality of drug perturbations* (emphasis added). Claim 1 has also been amended to clarify that each response profile results from a different drug perturbation *among said first plurality of drug perturbations* to said type of cell or organism. Claims 29 and 38 have been amended

similarly. Support for the amendments is found in the specification at page 41, lines 21-24 and page 43, lines 7-31.

Claim 18 has been amended to recite that in the claimed method the *cluster analysis is carried out by a hierarchical clustering method*; that step (a) involves determining *for each cluster generated by said cluster analysis* an actual fractional improvement in cluster analysis of the cellular constituents *based on the unpermuted responses of said cellular constituents*, step (d) involves determining *for each cluster generated in step (c)* the fractional improvement in the cluster analysis of cellular constituents *based on the permuted responses of cellular constituents*, step (e) involves repeating the steps of (b) through (d), i.e., generating permuted responses of cellular constituents, performing cluster analysis on the permuted responses of cellular constituents, and determining fractional improvement on the permuted data, so that a distribution of fractional improvements is obtained *for each cluster generated by said cluster analysis*; that in the claimed method the *fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters*; and that the statistical significance *for each of said sets of co-varying cellular constituents* is determined by comparing the actual fractional improvement *for the corresponding cluster* to the distribution of fractional improvements *for the corresponding cluster* (emphasis added). Claims 64, 96 and 123 have been amended similarly. Support for the amendments is found in the specification at page 28, line 28, through page 30, line 20. Claims 18, 64, 96 and 123 have also been amended to correct typographical errors.

Claim 26 has been amended to recite that in the claimed method the *cluster analysis is carried out by a hierarchical clustering method*; that step (a) involves determining *for each cluster generated by said cluster analysis* an actual fractional improvement in the cluster analysis of the response profiles, step (d) involves determining *for each cluster generated in step (c)* the fractional improvement in the cluster analysis on the permuted response profiles, step (e) involves repeating said steps of (b) through (d), i.e., generating permuted response profiles, performing cluster analysis on the permuted response profiles, and determining fractional improvement on the permuted data, so that a distribution of fractional improvements is obtained *for each cluster generated by said cluster analysis*; that in the claimed method the *fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters*; and that in the claimed method the

statistical significance *of each of said sets of response profiles* is determined by comparing the actual fractional improvement *for the corresponding cluster* to the distribution of fractional improvements *for the corresponding cluster* (emphasis added). Claims 50 and 106 have been amended similarly. Support for the amendments is found in the specification at page 28, line 28, through page 30, line 20; and at page 37, lines 17-23.

No new matter has been added by the amendments.

APPLICANTS' INTERVIEW SUMMARY

Applicants thank Primary Examiner Ardin Marschel for the courtesies extended during the telephone interview on January 8, 2002 (hereinafter "the Interview") with R. Douglas Bradley, Applicant Yudong He and Applicants' representatives Adriane M. Antler and Weining Wang. During the interview, the claim rejections under 35 U.S.C. § 112, first paragraph, 35 U.S.C. § 112, second paragraph, and 35 U.S.C. § 103(a) were discussed. The reference Eisen et al.,1998, *Proc. Natl. Acad. Sci. USA* 95:14863 was also discussed as it pertains to the claim rejections under 35 U.S.C. § 103(a).

The claim rejections in the instant Office Action under 35 U.S.C. § 112, first paragraph, were first discussed. The Examiner agreed that the printout of a review of S-plus entitled "S-plus in Teaching" by Henery faxed to the Examiner for review on January 7, 2002 is acceptable to demonstrate that both S-plus and *hclust* are well known in the art. The Examiner indicated that a submission of a total of three references published prior to the filing of the instant application by different groups would overcome the rejections. The Examiner also indicated that he would accept evidence showing that S-plus is still commercially available. Dr. Antler agreed to submit additional references and evidence of commercial availability of S-plus in the response to the Office Action.

The claim rejections in the instant Office Action under 35 U.S.C. § 112, second paragraph, were then discussed. Dr. Antler explained that the fractional improvement is computed with respect to a cluster in going from one cluster into two clusters in a clustering tree. Dr. Antler also propose to amend the claims to include such recitation. The Examiner indicated that he would reconsider the rejection if such a recitation is in included in the claims.

INTERVIEW SUMMARY OK MM 6-25-02

The interview participants then discussed the claim rejections under 35 U.S.C. § 103(a) over Eisen et al.,1998, *Proc. Natl. Acad. Sci. USA* 95:14863. Dr. Antler explained, as discussed below, that the reference does not make the claimed invention obvious. In particular, Eisen does not teach or suggest determining among a plurality of genesets, each of those that are upregulated or down regulated by a plurality of perturbations, and using these determined genesets as the consensus profile. Dr. Antler proposed to amend the claims to this effect. The Examiner agreed to consider such amendment.

The interview participants also discussed the claim rejections over claims reciting projection of cellular constituent sets, e.g., claim 29. Dr. Antler explained, as discussed below, that projection is independent of clustering, and thus Eisen's teaching of using supervised clustering to obtain clusters of genes teaches or suggests nothing about the projection of response profiles or projected response profiles. The Examiner agreed to reconsider the rejection.

The interview participants also discussed the claim rejections over claims reciting methods comprising a step of determining the statistical significance of obtained cellular constituent sets, e.g., claim 17. Dr. Antler explained, as discussed below, that Eisen does not teach or suggest a method comprising a step of determining the statistical significance of obtained cellular constituent sets. The Examiner agreed that Eisen does not teach or suggest a method comprising a step of determining the statistical significance of obtained cellular constituent sets and that the rejections of these claims will be withdrawn.

CORRECTION OF DRAWINGS

The Examiner has indicated that Applicants are required to submit drawing corrections within the time period set for responding to the Office Action. Applicants submit herewith formal drawings consisting of 15 sheets of drawings corresponding to Figures 1-11.

THE REJECTION UNDER 35 U.S.C. § 112, FIRST PARAGRAPH, SHOULD BE WITHDRAWN

Claims 14, 22, 47, 61, 92 and 119 are rejected under 35 U.S.C. § 112, first paragraph, as containing subject matter which was not described in the specification in such a way as to enable one skilled in the art to which it pertains, or with which it is most nearly connected, to make and/or use the invention. The Examiner contends that the algorithm of *hclust* is an

essential subject matter for the practice of the above-listed claims and as such cannot be enabled by incorporation by reference to a printed publication. Applicants respectfully disagree with the Examiner for the reasons set forth below.

A patent needs not teach, and preferably omits, what is well known in the art. *Spectra-Physics, Inc. v. Coherent, Inc.*, 827 F.2d 1524, 3 U.S.P.Q.2d 1737 (Fed. Cir. 1987). Applicants respectfully point out that software package S-Plus which includes *hclust* is a well known and widely used software package for performing statistical analysis and *hclust* is a well known algorithm for performing hierarchical cluster analysis. As evidence that S-plus and *hclust* algorithm are well known in the art, Applicants submit as Exhibit C a printout of a review of S-plus entitled "A Flavour of S-Plus" by Bowman and as Exhibit D a printout of a review of S-plus entitled "S-plus in Teaching" by Henery. Henery discloses that since 1989 S-plus was adopted as the official language for teaching all Statistics courses at University of Strathclyde, whereas Bowman discloses the use of S-plus as a teaching medium at University of Glasgow. Applicants also direct the Examiner's attention to Weinstein et al., 1997, Science 275:343-349, entitled "An information-intensive approach to the molecular pharmacology of cancer," already submitted as reference GO in the Information Disclosure Statement filed on October 5, 1999. As evidenced by note no. 21 on page 349, Weinstein uses S-plus in its cluster analysis calculations. Furthermore, Applicants submit as Exhibit E printouts of web pages of Insightful Corp., a vendor of S-plus, demonstrating that S-plus is currently commercially available. As can be seen from these references, S-plus and *hclust* are indeed both well known and widely used in the art. Therefore, anyone of skill in the art would be readily capable of performing the claimed method of cluster analysis of response profile data using the S-plus software and the *hclust* algorithm. Such a well known algorithm is preferably omitted in the specification. Therefore, Applicants respectfully submit that the rejection of claims 14, 22, 47, 61, 92 and 119 under 35 U.S.C. § 112, first paragraph, is in error, and should be withdrawn.

THE REJECTION UNDER 35 U.S.C. § 112, SECOND PARAGRAPH, SHOULD BE WITHDRAWN

Claims 18, 26, 50, 64, 96, 106 and 123 are rejected under 35 U.S.C. § 112, second paragraph, as allegedly being indefinite. The Examiner contends that claims 18, etc., are vague and indefinite regarding step (a) because in the step a fractional improvement is

determined but what is performed in order to obtain such an improvement is not indicated. In supporting his above-mentioned contention, the Examiner also contends that step (a) summarizes what is performed in steps (b)-(d).

Applicants have amended claims 18, 26, 50, 64, 96, 106 and 123 as described above. Applicants further respectfully point out that a fractional improvement is defined at page 29, lines 13-26, of the specification as an improvement in total scatter at a particular branch point in a cluster tree with respect to the cluster centers in going from one cluster to two clusters. Thus, a fractional improvement measures the change ("improvement") in total scatter if a cluster is split into two clusters at a branching point. Applicants have amended the claims to recite that in the claimed method the fractional improvement *is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters.* An actual fractional improvement is defined as the fractional improvement of the unpermuted data, i.e., data obtained from cluster analysis of the original data (see specification at page 29, lines 29-32). Thus, in the rejected claims, step (a) determines a fractional improvement of the unpermuted data; steps (b)-(d) perform permutation of the original data, cluster analysis on the permuted data, and determination of a fractional improvement based on the permuted data; and step (e) repeats the steps of (b)-(d), i.e., the steps of generating permuted responses of cellular constituents, performing cluster analysis on the permuted responses of cellular constituents, and determining fractional improvement based on the permuted data, to generate a distribution of fractional improvements. Therefore, Applicants respectfully submit that claims 18, 26, 50, 64, 96, 106 and 123 as amended are not indefinite, and that the rejection under 35 U.S.C. § 112, second paragraph, should be withdrawn.

THE REJECTIONS UNDER 35 U.S.C. § 103(a) SHOULD BE WITHDRAWN

Claims 1, 3-8, 10-13, 15-17, 19-21, 23-25, 27-46, 48, 49, 58-60, 62, 63, 72-78, 89-91, 93-95, 97-100, 105, 107-113, 115-118, 120-122 and 124 are rejected under 35 U.S.C. § 103(a) as being unpatentable over Eisen et al.,1998, *Proc. Natl. Acad. Sci. USA* 95:14863 ("Eisen"). The Examiner contends that "it would have been obvious to someone of ordinary skill in the art at the time of the instant invention to perform the genome-scale expression analysis of the reference with the clustering of data in order to determine those sets of genes which are affected as to expression by various conditions." The Examiner also contends that

"the usage of supervised clustering in the reference as a known reference vector is reasonably interpreted as the projected profiles of instant claim 29, for example, when further utilized in subsequent analyses as suggested in the references." Claims 1, 3-8, 10-13, 15-17, 19-21, 23-25, 27-46, 48, 49, 58-60, 62, 63, 72-78, 89-91, 93-95, 97-100, 105, 107-118, 120-122 and 124 are rejected under 35 U.S.C. § 103(a) as being unpatentable over Eisen in view of Welsh, U.S. Patent No. 5,686,114 ("Welsh"). The Examiner contends that "Welsh suggests and motivates the issue of the study of drug toxicity along with dosing or treatment in drug targeting .... it would have been obvious to someone of ordinary skill in the art at the time of the instant invention to also profile drug toxicity along with efficacy in evaluating profiles of drug perturbations as instantly claimed." Applicants respectfully disagree with the Examiner for reasons set forth below.

A finding of obviousness under 35 U.S.C. § 103(a) requires a determination that the differences between the claimed subject matter and the prior art are such that the subject matter as a whole would have been obvious to one of ordinary skill in the art at the time the invention was made. *Graham v. Deere*, 383, U.S. 1 (1956). The relevant inquiry is whether the prior art suggests the invention and whether the prior art provides one of ordinary skill in the art with a reasonable expectation of success. Both the suggestion and the reasonable expectation of success must be found in the prior art. *In re Vaeck*, 947 F.2d 488 (Fed. Cir. 1991).

Eisen teaches cluster analysis for analyzing the genome-wide expression data obtained from DNA microarray measurements. In Eisen, a microarray containing essentially every ORF from yeast is used to measure gene expression data of budding yeast during the diauxic shift, the mitotic cell division cycle, sporulation, and temperature and reducing shocks and a microarray with 9,800 cDNAs representing 8,600 distinct human transcripts is used to measure gene expression data of primary human fibroblasts stimulated with serum following serum starvation. The gene expression data obtained from the microarray measurement are then analyzed using cluster analysis to identify gene expression patterns. Eisen suggests that genes of similar function cluster together. However, although Eisen teaches that *genes* can co-vary and therefore can be clustered into co-varying sets, Eisen does not teach or suggest that some of such co-varying *sets of genes or clusters of genes*, i.e., a group of co-varying sets of genes, can be upregulated or downregulated by a particular collection of different drug

perturbations. Applicants note that genes clustered in different sets are not co-varying in general (that is why they are clustered in different clusters) and do not in general respond to different perturbations similarly. Nor does Eisen teach or suggest that the responses of the *sets of genes* that similarly respond to a particular collection of different drug perturbations can be used as the consensus profile for representing the response profiles of a cell in response to such a collection of drug perturbations. Eisen's teaching that genes can be clustered into co-varying sets does not provide one of ordinary skill in the art with a suggestion and reasonable expectation of success that a group of the co-varying sets can have similar responses, i.e., upregulated or downregulated, to a collection of different drug perturbations and that such a group of the co-varying sets of cellular constituents can be identified and their responses used to represent the response of a cell to the collection of drug perturbations, e.g., as the consensus profile of the cell in response to the collection of drug perturbations. Thus, Eisen does not teach a method comprising *determining, for each of* a plurality of sets of cellular constituents in a plurality of response profiles, *whether said set of* cellular constituents is upregulated or downregulated by said first plurality of drug perturbations, each response profile in said plurality of response profiles (i) comprising measurements of a plurality of cellular constituents, and (ii) resulting from a different drug perturbation *among said first plurality of drug perturbations* to said type of cell or organism, wherein each set of cellular constituents in said plurality of sets of cellular constituents consists of cellular constituents that co-vary under a second plurality of perturbations or that are co-regulated, wherein said plurality of response profiles comprises at least five response profiles, and wherein *said consensus profile for said first plurality of drug perturbations comprises measurements of said set or sets of cellular constituents that are determined in said determining step to be upregulated or downregulated by said first plurality of drug perturbations*.

The Examiner also contends that the usage of supervised clustering in Eisen is reasonably interpreted as the projected profiles of instant claims, e.g., claim 29. Applicants respectfully submit that the Examiner's contention is erroneous. At the outset, Applicants respectfully point out that the projection of response profiles according to a definition of cellular constituent sets is independent as to how the cellular constituent sets are defined and obtained. Projection of response profiles is carried out *after* the cellular constituent sets have

been obtained by, e.g., either supervised or unsupervised clustering. The difference between unsupervised and supervised clustering is whether predefined reference vectors are used to obtain the cellular constituent sets. In contrast, as described in Section 5.3.4. of the specification, projection of response profiles involves representing the response profiles in terms of the basis cellular constituent sets, which are obtained from the cellular constituent sets (see Section 5.3.2. for a description of basis cellular constituent sets). Therefore, Eisen's teaching of using supervised clustering to obtain clusters of genes teaches or suggests nothing about the projection of response profiles or projected response profiles.

In addition, regarding claims 17, 25, 49, 63, 95, and 122, Eisen does not teach or suggest a method comprising a step of determining the statistical significance of obtained cellular constituent sets.

Welsh teaches pharmaceutical compositions comprising an inorganic pyrophosphate for use in the treatment of a disease associated with inappropriate or inadequate ATP-binding cassette (ABC) protein activity. Welsh does not teach or suggest what is missing in Eisen, i.e., that a group of the co-varying cellular constituent sets can have similar responses, i.e., upregulated or downregulated, to a collection of different drug perturbations and that such a group of the co-varying sets of cellular constituents can be identified and their responses used to represent the response of a cell to the collection of drug perturbations, e.g., as the consensus profile of the cell in response to the collection of drug perturbations. Welsh does not teach or suggest projection of response profiles onto basis cellular constituent sets or projected response profiles. Nor does Welsh teach or suggest a method comprising a step of determining the statistical significance of obtained cellular constituent sets.

Therefore, Applicants respectfully submit that neither Eisen nor Eisen in view of Welsh renders the present invention unpatentable, and that the rejection under 35 U.S.C. § 103(a) over Eisen and the rejection under 35 U.S.C. § 103(a) over Eisen in view of Welsh should be withdrawn.

CONCLUSION

Applicants respectfully request entry of the foregoing amendments and remarks into the file of the above-identified application. Applicants believe that each ground for rejection has been successfully overcome or obviated, and that all the pending claims are in condition

for allowance. Withdrawal of the Examiner's rejections and allowance of the application are respectfully requested.

Respectfully submitted,

Date April 9, 2002

Adriane M. Antler 32,605
Adriane M. Antler (Reg. No.)

PENNIE & EDMONDS LLP
1155 Avenue of the Americas
New York, New York 10036-2711
(212) 790-9090

Enclosures

EXHIBIT A: MARKED VERSION OF THE AMENDED CLAIMS
U.S. APPLICATION SERIAL NO. 09/220,142
(ATTORNEY DOCKET NO. 9301-035-999)

(as amended April 9, 2002)

1. (Five Times Amended) A method of determining a consensus profile for a first plurality of drug perturbations to a cell type or organism, said method comprising [identifying among] determining, for each of a plurality of sets of cellular constituents in a plurality of response profiles [one or more sets of cellular constituents], [each of said one or more sets of] whether said set of cellular constituents [being] is upregulated or downregulated by said first plurality of drug perturbations, each response profile in said plurality of response profiles (i) comprising measurements of a plurality of cellular constituents, and (ii) resulting from a different drug perturbation among said first plurality of drug perturbations to said type of cell or organism, wherein each set of cellular constituents in said plurality of sets of cellular constituents consists of cellular constituents that co-vary under a second plurality of perturbations or that are co-regulated, wherein said plurality of response profiles comprises at least five response profiles, and wherein said consensus profile for said first plurality of drug perturbations comprises measurements of said [one or more] set or sets of cellular constituents that are determined in said determining step to be upregulated or downregulated by said first plurality of drug perturbations.

18. (Twice Amended) The method of claim 17, wherein said cluster analysis is carried out by a hierarchical clustering method, and wherein the objective statistical test comprises:

(a) determining for each cluster generated by said cluster analysis an actual fractional improvement in the cluster analysis of the cellular constituents based on the unpermuted responses of said cellular constituents, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters;

(b) generating permuted [response] responses of cellular constituents by means of Monte Carlo randomization of perturbation index for the response of each cellular constituent across all perturbations;

(c) performing said cluster analysis on the permuted [response] responses of cellular constituents;

(d) determining for each cluster generated in step (c) the fractional improvement in the cluster analysis of cellular constituents based on the permuted [response] responses of cellular constituents, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters; and

(e) repeating [said] steps [of generating permuted response of cellular constituents and performing cluster analysis on the permuted response of cellular constituents] (b) through (d) so that a distribution of fractional improvements in the cluster analysis of the cellular constituents is obtained for each said cluster generated by said cluster analysis;

wherein the statistical significance of each of said sets of co-varying cellular constituents is determined by comparing the actual fractional improvement for the corresponding cluster to the distribution of fractional improvements for the corresponding cluster.

26. (Twice Amended) The method of claim 25, wherein said cluster analysis is carried out by a hierarchical clustering method, and wherein the objective statistical test comprises:

(a) determining for each cluster generated by said cluster analysis an actual fractional improvement in the cluster analysis of the unpermuted response profiles, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters;

(b) generating permuted response profiles by means of Monte Carlo randomization of cellular constituent index for each response profile across the measured cellular constituents;

(c) performing said cluster analysis on the permuted response profiles;

(d) determining for each cluster generated in step (c) the fractional improvement in the cluster analysis [on] of the permuted response profiles, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters; and

(e) repeating [said] steps [of generating permuted response profiles and performing cluster analysis on the permuted response profiles] (b) through (d) so that a distribution of fractional improvements in the cluster analysis of the response profiles is obtained for each said cluster generated by said cluster analysis;

wherein the statistical significance of each of said sets of response profiles is determined by comparing the actual fractional improvement for the corresponding cluster to the distribution of fractional improvements for the corresponding cluster.

29. (Three Time Amended) A method of determining a consensus profile for a first plurality of perturbations to a cell type or organism, said method comprising [identifying among] determining, for each of a plurality of sets of cellular constituents in a plurality of projected profiles [one or more sets of cellular constituents], [each of said one or more sets of] whether said set of cellular constituents [being] is upregulated or downregulated by said first plurality of perturbations, each projected profile in said plurality of projected profiles

(i) resulting from a different perturbation among said first plurality of perturbations to said type of cell or organism, and

(ii) comprising measurements of a plurality of cellular constituents in said type of cell or organism that have been projected onto basis cellular constituent sets, said basis cellular constituent sets being defined by co-variation of measurements of cellular constituents under a second plurality of different perturbations, wherein said consensus profile for said first plurality of perturbations comprises projected measurements of said [one or more] set or sets of cellular constituents that are determined in said determining step to be upregulated or downregulated by said first plurality of perturbations.

38. (Four Times Amended) A method of determining a consensus profile for a first plurality of perturbations to a cell type or organism, said method comprising [identifying among] determining, for each of a plurality of sets of genes in a plurality of response profiles [one or more sets of genes], [each of said one or more sets of] whether said set of genes [being] upregulated or downregulated by said first plurality of perturbations, each response profile in said plurality of response profiles (i) comprising measurements of transcript levels

for a plurality of genes, and (ii) resulting from a different perturbation among said first plurality of perturbations to said type of cell or organism, wherein each set of genes in said plurality of sets of genes consists of genes having transcripts that co-vary under a second plurality of perturbations or that are co-regulated, and wherein said consensus profile for said first plurality of perturbations comprises measurements of transcript levels for said [one or more] set or sets of genes that are determined in said determining step to be upregulated or downregulated by said first plurality of perturbations.

50. (Twice Amended) The method of claim 49, wherein said cluster analysis is carried out by a hierarchical clustering method, and wherein the objective statistical test comprises:

(a) determining for each cluster generated by said cluster analysis an actual fractional improvement in the cluster analysis of the unpermuted response profiles, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters;

(b) generating permuted response profiles by means of Monte Carlo randomization of gene index for each response profile across the measured genes;

(c) performing said cluster analysis on the permuted response profiles;

(d) determining for each cluster generated in step (c) the fractional improvement in the cluster analysis of the permuted response profiles, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters; and

(e) repeating [said] steps [of generating permuted response profiles and performing cluster analysis on the permuted response profiles] (b) through (d) so that a distribution of fractional improvements in the cluster analysis of the response profiles is obtained for each cluster generated by said cluster analysis;

wherein the statistical significance of each of said sets of response profiles is determined by comparing the actual fractional improvement for the corresponding cluster to the distribution of fractional improvements for the corresponding cluster.

64. (Twice Amended) The method of claim 63, wherein said cluster analysis is carried out by a hierarchical clustering method, and wherein the objective statistical test comprises:

(a) determining for each cluster generated by said cluster analysis an actual fractional improvement in the cluster analysis of cellular constituents based on the unpermuted responses of said cellular constituents, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters;

(b) generating permuted [response] responses of cellular constituents by means of Monte Carlo randomization of the perturbation index for each cellular constituent across all perturbations;

(c) performing said cluster analysis on the permuted [response] responses of cellular constituents;

(d) determining for each cluster generated in step (c) the fractional improvement in the cluster analysis of cellular constituents based on the permuted [response] responses of cellular constituents, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters; and

(e) repeating [said] steps [of generating permuted response of cellular constituents and performing cluster analysis on the permuted response of cellular constituents] (b) through (d) so that a distribution of fractional improvements in the cluster analysis of the cellular constituents is obtained for each cluster generated by said cluster analysis;

wherein the statistical significance of each of said sets of cellular constituents is determined by comparing the actual fractional improvement for the corresponding cluster to the distribution of fractional improvements for the corresponding cluster.

96. (Amended) The method of claim 95, wherein said cluster analysis is carried out by a hierarchical clustering method, and wherein the objective statistical test comprises:

(a) determining for each cluster generated by said cluster analysis an actual fractional improvement in the cluster analysis of the cellular constituents based on the unpermuted responses of said cellular constituents, wherein said

fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters;

(b) generating permuted [response] responses of cellular constituents by means of Monte Carlo randomization of the perturbation index for response of each cellular constituent across the set of perturbations;

(c) performing said cluster analysis on the permuted [response] responses of cellular constituents;

(d) determining for each cluster generated in step (c) the fractional improvement in the cluster analysis of cellular constituents based on the permuted response responses of cellular constituents, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters; and

(e) repeating [said] steps [of generating permuted response of cellular constituents and performing cluster analysis on the permuted response of cellular constituents] (b) through (d) so that a distribution of fractional improvements in the cluster analysis of the cellular constituents is obtained for each cluster generated by said cluster analysis,

wherein the statistical significance of each of said sets of co-varying cellular constituents is determined by comparing the actual fractional improvement for the corresponding cluster to the distribution of fractional improvements for the corresponding cluster.

106. (Amended) The method of claim 105, wherein said cluster analysis is carried out by a hierarchical clustering method, and wherein the objective statistical test comprises:

(a) determining for each cluster generated by said cluster analysis an actual fractional improvement in the cluster analysis of the unpermuted response profiles, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters;

(b) generating permuted response profiles by means of Monte Carlo randomization of cellular constituent index for each response profile across the measured cellular constituents;

(c) performing said cluster analysis on the permuted response profiles;

(d) determining for each cluster generated in step (c) the fractional improvement in the cluster analysis of the permuted response profiles, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters; and

(e) repeating [said] steps [of generating permuted response profiles and performing cluster analysis on the permuted response profiles] (b) through (d) so that a distribution of fractional improvements in the cluster analysis of the response profiles is obtained for each cluster generated by said cluster analysis;

wherein the statistical significance of each of said sets of response profiles is determined by comparing the actual fractional improvement for the corresponding cluster to the distribution of fractional improvements for the corresponding cluster.

123. (Amended) The method of claim 122, wherein said cluster analysis is carried out by a hierarchical clustering method, and wherein the objective statistical test comprises:

(a) determining for each cluster generated by said cluster analysis an actual fractional improvement in the cluster analysis of the genes based on the unpermuted responses of said genes, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters;

(b) generating permuted [response] responses of genes by means of Monte Carlo randomization of perturbation index for the response of each gene across all perturbations;

(c) performing cluster analysis on the permuted [response] responses of genes;

(d) determining for each cluster generated in step (c) the fractional improvement in the cluster analysis of genes based on the permuted [response] responses of genes, wherein said fractional improvement is an improvement in total scatter with respect to a cluster center in going from one cluster to two clusters; and

(e) repeating [said] steps [of generating permuted response of genes and performing cluster analysis on the permuted response of genes] (b) through (d)

so that a distribution of fractional improvements in the cluster analysis of the genes is obtained for each cluster generated by said cluster analysis;

wherein the statistical significance of each of said sets of co-varying genes is determined by comparing the actual fractional improvement for the corresponding cluster to the distribution of fractional improvements for the corresponding cluster.

cti STATISTICS

A Flavour of S-Plus

by Adrian Bowman
University of Glasgow

This article appears in the February 1993 issue of the newsletter ***Maths&Stats****, as part of a special S-Plus supplement. It is based on a talk given at the S-Plus Workshops in 1992.*

Figures

In introducing any new package, it is probably easiest to contrast it with a package which is already well known, such as MINITAB which is widely used in teaching Introductory Courses throughout the country. Something of the style of S-Plus is indicated by the following statements.

```
x < - scan ("haem.dat")
d < - matrix (x, ncol = 6, byrow = T)
pairs (d)
```

The effect of the first two of these statements is to read a file of data, referring to blood measurements of a group of workers and reported by Royston (Applied Statistics 1983, 32, 121-33) in a matrix called d. Afficionados of MINITAB might instantly claim that this is much more cumbersome than MINITAB's simple 'read' command. The first S-Plus statement stores all the data as a long string of numbers in the vector x. The second statement then reorganises this by filling up a matrix d in seven columns along the rows. One advantage of this is the flexibility of being able to handle files where records are spread perhaps irregularly over several lines of a file. These commands also indicate another feature of S-Plus which is that everything is done through the use of functions. The assignment operator <- takes the result of the function and stores it in the object on the left hand side. The functional nature of the language means that statements can be combined as, for example,

```
d < - matrix (scan ("haem.dat"), ncol = 6, byrow = T)
```

The third original statement, pairs(d), illustrates one of the main strengths of S-Plus. The

result of this function is to take the columns of the matrix d and produce all the pair-wise scatterplots which can be created from these columns. The resulting picture is displayed in Figure 1. S-Plus is extremely good at producing high quality graphics which, given the right equipment, can be sent to a laser printer at the press of an electronic button.

As a further example of this, the following command produces a dendrogram as displayed in Figure 2:

```
plclust (hclust (dist(d), method = "connected"))
```

Here the parameter method = "connected" selects the use of single link in clustering. After using packages where the standard output from a cluster analysis is a long list of numerical information, this form of high quality graphics is a delight. S-Plus is particularly strong in multivariate methods and has routines for a wide variety of techniques. Most of these graphical methods are extremely easy to use. As a package for giving students relatively painless experience of more advanced techniques, S-Plus is an extremely useful teaching medium.

The role of packages in allowing students to carry out data analysis and to provide an environment within which to model data is probably the most important one. There are however other uses of the computer in teaching. One of these is to take basic ideas or techniques and allow students to explore the meaning and properties of them. This is where the fact that S-Plus is a full blooded programming language comes into its own. It is very easy to string together basic S-Plus commands and to create these as a new function. Figure 3 illustrates the output of a function designed to take 100 samples of size 25 from a normal population with mean 69 and display the resulting computed confidence intervals. This simple graphical illustration communicates the fact that confidence intervals are random intervals and that confidence comes from the fact that, on average, 95% of these intervals will capture the true value. Simulations can be repeated by calling the function as many times as required. In the past it has been necessary to write special purpose software to produce this kind of graphical illustration but the presence of packages with sophisticated graphics and programming facilities now mean that we have the advantage of being able to do this kind of exercise within an environment which has all the standard tools at our disposal.

Figure 4 gives some output from a function written to display repeated measurements data. With this type of data observations are repeated on individuals across time. The data here refer to the levels of leutimizing hormone in two groups of cows, reported by Raz (Biometrics (1989) 54, 851-71). It is a very commonly occurring data structure but one which many data packages find difficult to handle. Some years ago, I wrote a programme in BBCBasic on a pc which took up several pages of code. The function to produce these repeated measurements plot takes up less than two pages of S-Plus code and is far more flexible and powerful.

The acid test of any particular package is whether it is used in practice. I can say that I am now a wholehearted and enthusiastic S-Plus user although it is not yet available within my own department as a teaching vehicle. I use it routinely for research and consulting problems. As previously indicated, its strengths lie in the high quality nature of the graphics, the easy access to some sophisticated and powerful modelling tools, and the flexible nature of the programming environment.

With any package there will always be disadvantages. Until recently there were some surprising omissions in the area of elementary techniques and of designed experiments. This has to a large extent been remedied in version 3 of the package. The documentation is not yet ideal as the result of the history of the development of the software but this is also improving. One area which is still handled rather patchily is that of missing data. Some functions will cope with this whereas others will not operate when missing data are detected. Despite these disadvantages, S-Plus is an extremely powerful and sophisticated package. It was designed for UNIX systems and works extremely well on Sun workstations, although it can be slow if 'for' loops are employed. It is now available for DOS too, and John Hinde has reviewed this in an article.

Adrian Bowman
adrian@stats.gla.ac.uk

Figure 1

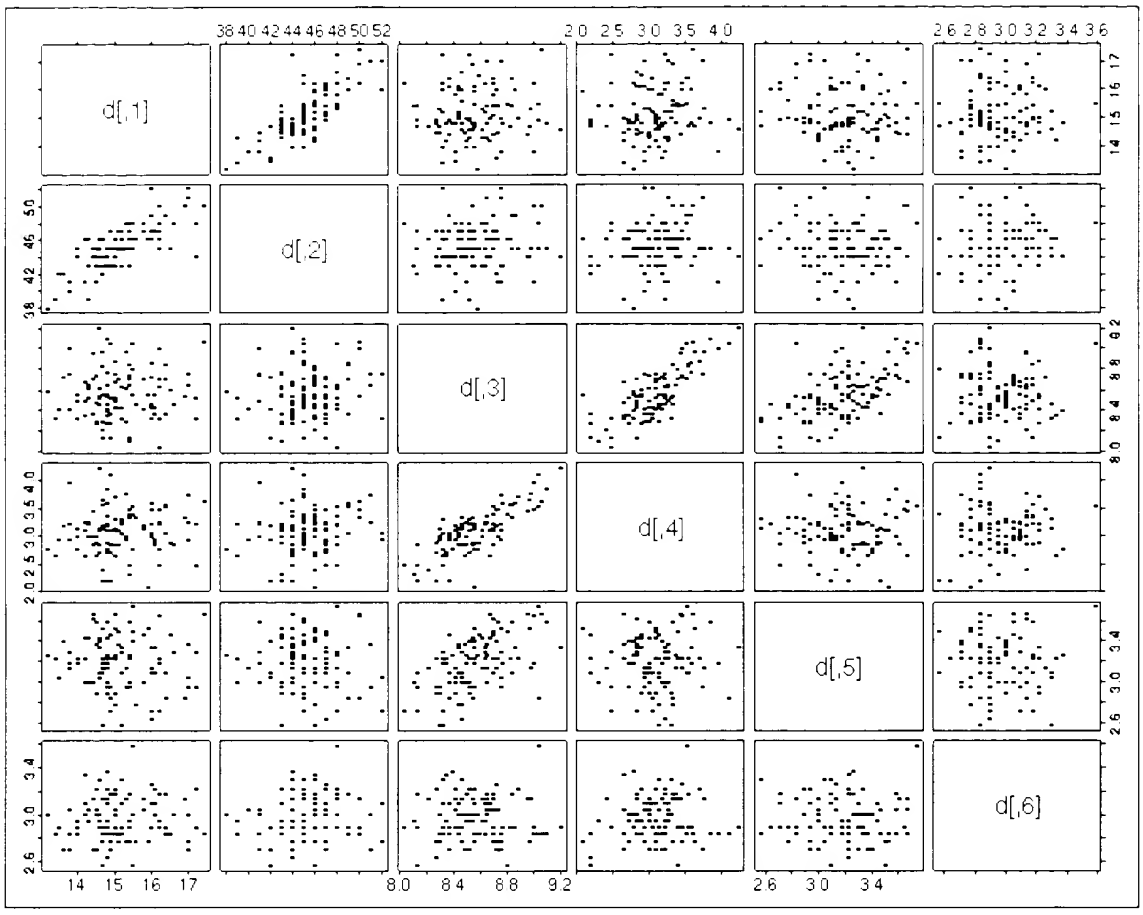

d[,1]
d[,2]
d[,3]
d[,4]
d[,5]
d[,6]
38 40 42 44 46 48 50 52
20 25 30 35 40
26 28 30 32 34 36
14 15 16 17
38 42 46 50
8.0 8.4 8.8 9.2
20 25 30 35 40
2.6 3.0 3.4
2.6 3.0 3.4
14 15 16 17
8.0 8.4 8.8 9.2
26 30 34

Figure 2

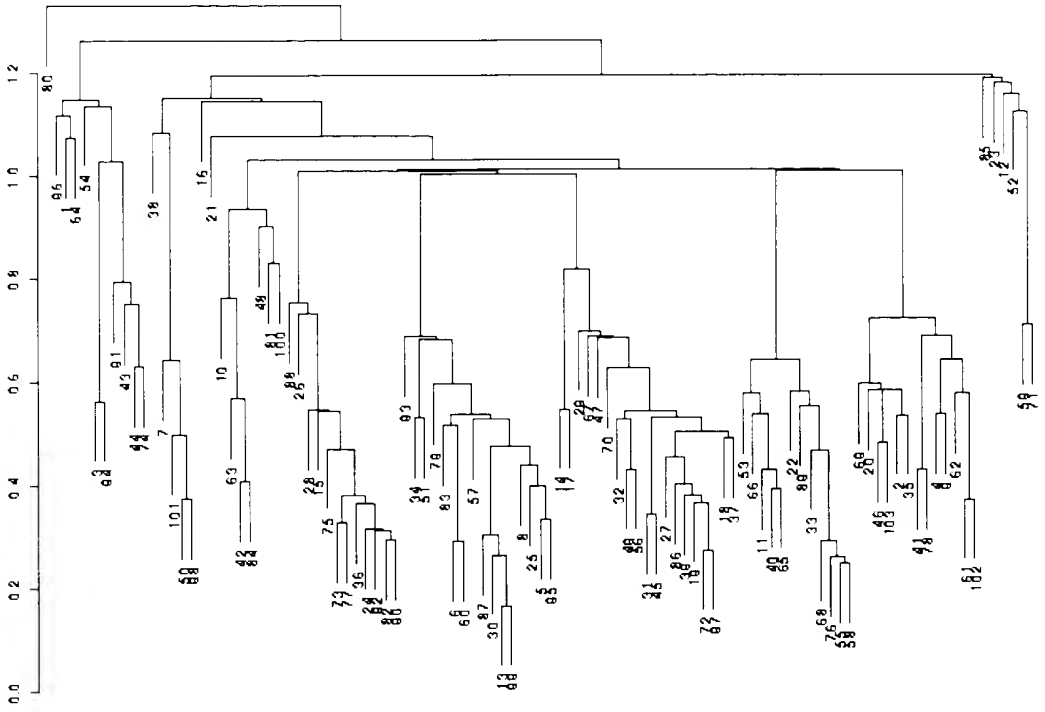


Figure 3

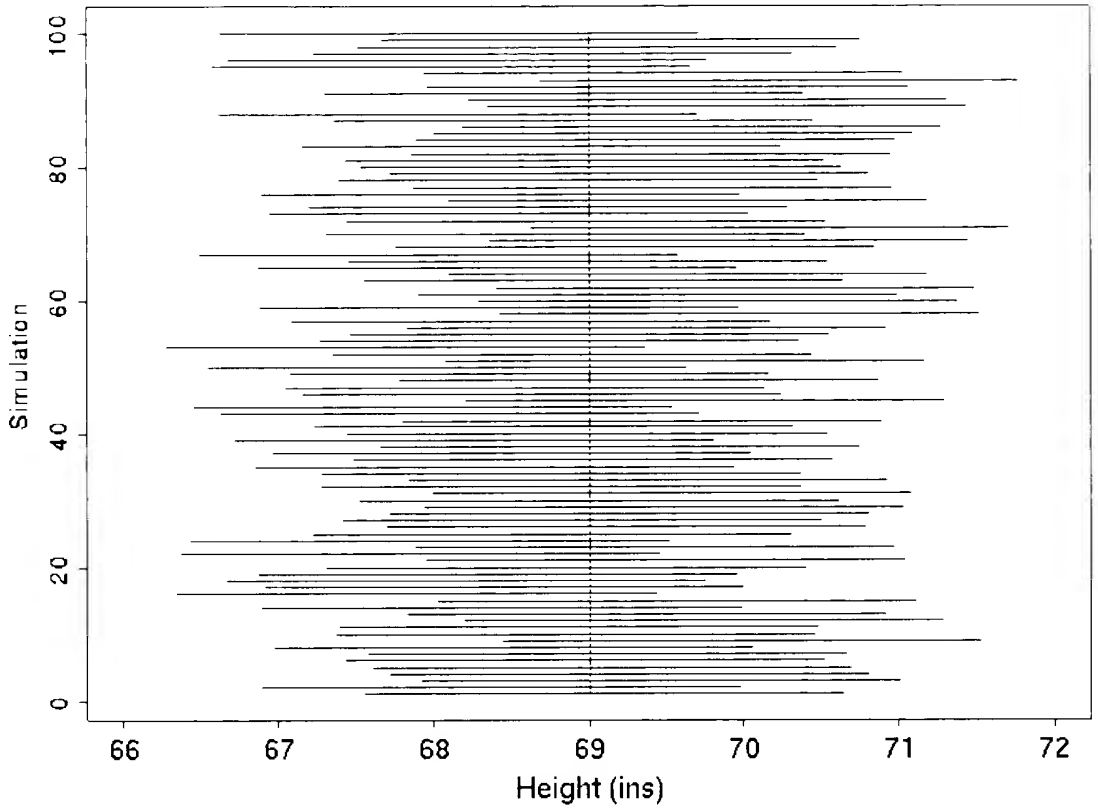
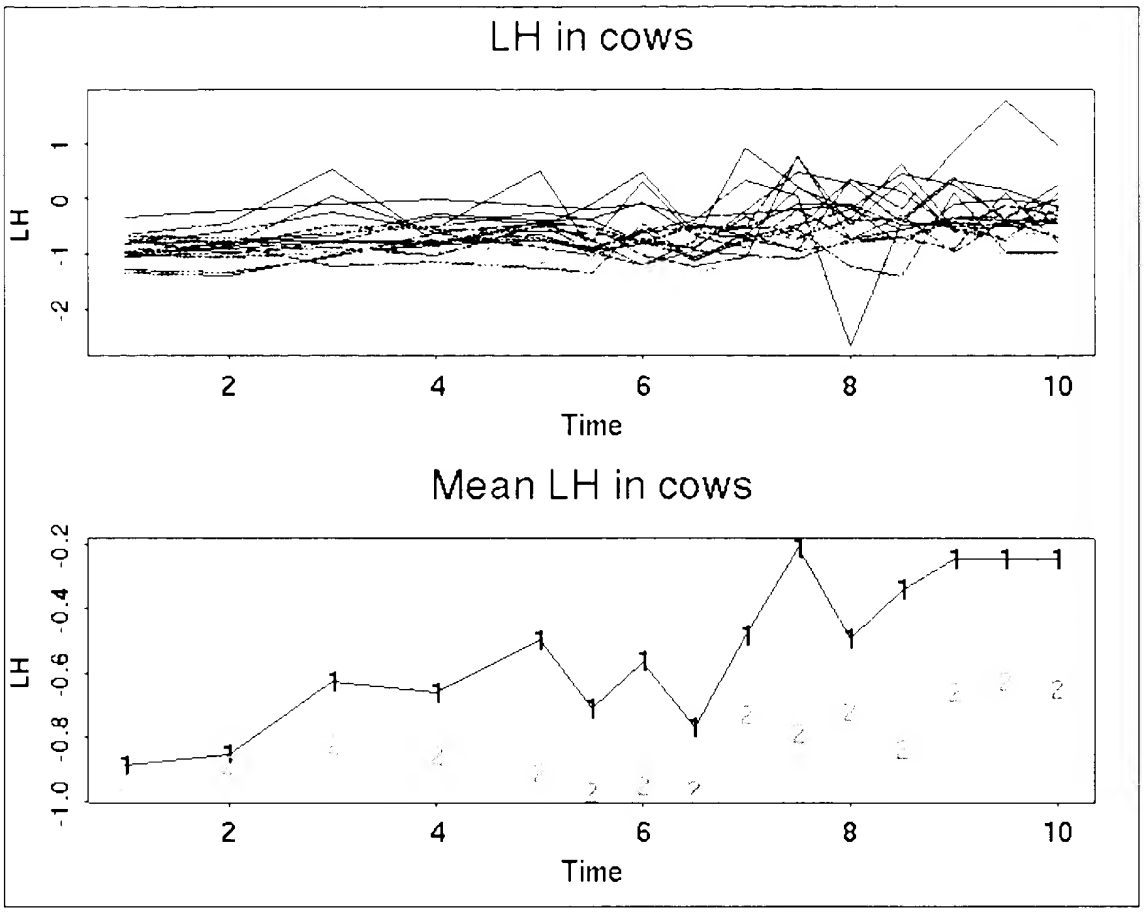


Figure 4


LH in cows
LH
1
0
-1
-2
2
4
6
8
10
Time
Mean LH in cows
LH
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
Time

cti STATISTICS

S-Plus in Teaching

by Bob Henery
University of Strathclyde

This article appears in the February 1993 issue of the newsletter ***Maths&Stats****, as part of a special S-Plus supplement. It is based on a talk given at the S-Plus Workshops in 1992.*

Background

For several years, I had been giving a 20-hour course to Final Year Honours students at Strathclyde University. The course covered non-normal data such as contingency tables, Generalised Linear Models, canonical variates, etc. with the computing side based on MINITAB and GLIM. At least half of the time was spent on theory, partly because no suitable introductory text was available for many years, although the situation has improved with the appearance of the second edition of McCullagh and Nelder (1990).

However, with the installation of a laboratory of 20 Sun workstations in October 1989, S-plus was adopted as the 'official' language for teaching all Statistics courses, (not only Final Year), and this required a complete revision of the content and format of the course. To make best use of the strong points of S-plus, modern statistical methods combined with excellent graphics, it was decided to give the minimum of theory and to concentrate on demonstrations, practicals and commentaries on the examples provided. The advantage was that the students could get some experience in using a large number procedures, the disadvantage being that each topic was covered superficially.

Developments in Splus

Splus is a modern statistical package: it is constantly being improved by the addition of new procedures and its basic structure also undergoes an evolutionary process. Each new version

brings with it a greater range of functions and greater power in the ability to process classes of datasets. Improvements in performance are not always monotonic however, and Splus version 3.0 was noticeably slower than Splus 2.3, a situation that was rapidly corrected with the introduction in October 1992 of version 3.1 which is best of all models to date. Make sure therefore, when ordering Splus, that the version number is 3.1 or later.

Originally the course was based on Splus 2.2 (October 1989), but that was updated to Splus 2.3 (June 1990) and thereafter to Splus 3.0 (June 1992). It will be necessary to update to version 3.1 soon. Each version entails some re-learning of the system for teacher and students alike.

Documentation

There are now three text-books dealing with the language S:

1. "The New S Language" by Becker, R., Chambers J.M. and Wilks A.R. (1988).;
2. "Statistical Models in S", eds. Chambers J.M. and Hastie, T.J. (1992);
3. "Data Analysis by Using S" by Sibuya, M. and Shibata, R. (1992).

The first is more concerned with programming; the second deals extensively with a few selected statistical models and their treatment in S; and the third is aimed more at students.

The S-PLUS User's Manual is also a very useful guide to statistical and graphical procedures, but it is probably not suitable for students who will want to use the online help facility on a narrow range of facilities.

Help - functions and data

Online information is available on both procedures and datasets using the help() command. The command help(plclust), for example, creates a window containing a complete specification of the plclust procedure, just as it appears in the New-S book. Of particular importance for teaching purposes is the liberal use of examples in the help files, which may be copied directly to the student's command window, and executed without further ado since they mostly refer to the provided datasets, of which there are several. Another good feature is that the help documentation contains very informative summaries of the theory and applications of important statistical procedures. By sending appropriate help files to the laser printer, each student may prepare for himself a customised set of notes, incuding worked examples. Another useful feature is that brief descriptions of the datasets are also given, for example on Fisher's iris data by help(iris).

Data Manipulation

Most S is done by assigning expressions. Objects in S may be of type logical, numeric, character (alphanumeric), and structures may be vectors, matrices, or lists. To give some feel for how S looks in practice, here is a complete example (given in help(plclust)) for generating a hierarchical clustering plot.

```
# the example plot is produced by:
suntools()
sums <- apply(author.count,1,sum)
adjusted <- sweep(author.count,1,sums,"/")
par(mar=c(18,4,4,1))
plclust(hclust(dist(adjusted)),label=dimnames(author.count)[[1]])
title("Clustering of Books Based on Letter Frequency")
```

Even complete novices can run the above example by clicking the mouse buttons and copying and pasting. It only remains to motivate the procedure (hierarchical clustering) and to give a commentary on the output, the graphical display of which is shown as Figure 1. Incidentally, the command used to add the caption "Figure 1" to the graph is

```
text(locator(1), "Figure 1",cex=2.0)
```

the caption being superimposed at the cursor when the mouse is clicked. Unfortunately, what you see on the graph is not always what you get on the laser printer. However, this difficulty is readily overcome by trial and error, and with almost no effort at all students are producing high-quality graphics for their reports.

The dimnames function used in the above example is another useful feature. Names of variables or objects (in the above example, names of books and authors) may be used to label graphs or to identify outliers. Dimnames are propagated through many procedures, with the result that residuals can be identified readily with the associated object or variable.

Graphical Procedures

These are excellent, with a range of modern procedures, such as brush and spin (which students enjoy). If required full control may be exercised over the format and content of all graphs, by labelling, adding lines or text etc. Once completed, graphs may be sent to a laser printer or Hewlett-Packard plotter.

Interfaces to C/FORTRAN

Interfaces to C and FORTRAN are possible, but for teaching purposes these are best hidden. However, course organisers who find that their favourite procedure is not supplied, may wish to code it themselves in FORTRAN say, and it may then be interfaced to S-plus. There are several restrictions on such FORTRAN routines, so some rewriting of procedures is almost certain. More probably, any deficiencies may be rectified by writing new procedures in S.

Statistical Procedures

There are three sources for S-plus procedures. Standard S-plus procedures are provided by Statistical Sciences Inc. who also give user support for these procedures. Because the standard S-plus functions do not include simple tests, a suite of programs called NESI (New Environment for Statistical Inference) have been provided by Prof. Shibata and colleagues at Keio University, Japan. The NESI functions are now included in the Splus package. Finally, the odd user- supplied function may be spotted in the mailing list *s-news* or at Statlib (see the Statistics Resources on the Web pages for more information).

S-plus has far too many functions to list, so here is a representative list of functions new to that version of Splus:

```
S-plus 2.2      hclust - hierarchical clustering
                mstree - minimum spanning tree
                discr  - discrimination
                prcomp - principal components

S-plus 2.3      glim   - generalised linear models
                ace    - alternating conditional expectation
                avas   - additive nonlinear regression with variance
                                stabilization
                ppreg  - projection pursuit regression
                nlmin  - non-linear minimization

S-plus 3.0      glm       - generalised linear models with factors,
                                interactions, etc.
                var.test - F variance ratio test
                t.test    - one or two sample t test
                wilcox.test - Wilcoxon signed rank or Mann-Whitney test
```

Teaching

From the outset, the approach was to introduce the students to as many important modern techniques as possible. S-plus is strong on exploratory data analysis, so at times the theoretical treatment was at best sketchy, and even non-existent. However, by going over to 100% course- work assessment, a modular approach was possible by which each topic could be taken in isolation. Each topic, which might involve two or three related S-plus procedures, was dealt with by some introductory lectures and demonstrations followed by a course-work assessment involving a min-project plus report, the whole topic occupying two weeks. In writing up the report, the emphasis was on the procedures in question and not on the particular dataset, and a full analysis of the data was not required. In this way many more procedures were discussed than formerly, and although the drawback is that the treatment was more superficial, yet the students enjoyed the course more, and felt that it was useful to know about as many procedures as possible. However, some students felt that more lectures

would have been useful.

Some of the newer S-plus procedures will be unfamiliar to the lecturers, never mind the students, and some have not stood the test of time. So, as an experiment, a couple of procedures were introduced with no preparatory lectures, the idea being to see how the students would cope with no aid from the lecturer, relying solely on their natural intelligence and the help facility. This experiment was only moderately successful, (was the fault in their intelligence or the help facility?), but I will try again next year!

Library

Collections of procedures may be gathered together for a specific course or a suite of programs for a specific purpose may conveniently be placed in a library. For example, Brian Yandell has developed a suite of programs under the heading 'penalised likelihood generalised linear models', and these were put in a library(pglm). Procedures for my own course, which included a customised version of discr for multivariate analysis of variance and a specially written procedure for correspondence analysis, were placed in library(da2). Also in library(da2) were instructions for projects, and the writing of reports etc.

With Splus version 3.0 and later it will be particularly easy to construct a suite of procedures dealing with classes of datasets.

Equipment

The teaching lab for the course now consists of 20 Sun Sparcstations 1+: the older 3/80 workstations were just a little on the slow side. The workstations are served by a central Sun 386i, and have access to a QMS810 laser printer and Hewlett-Packard 7550A graph plotter. Typically there are 15-20 students in the class, and occasionally there is a log-jam at the printer when graphs and help files are in great demand, but a modicum of discipline smooths the problem away.

S-plus requires a Unix environment, and this means that a workstation is advisable. Although a PC version is now available, I believe it requires a considerably beefed-up PC with lots of memory and a fast disc, and I do not know the relative merits of the Sun and PC versions.

Comparison with MINITAB, GLIM, ...

The S-plus package has been reviewed by D.G. Fraser, who compares the scope and complexity of S-plus with SAS, BMDP or Genstat (in Bull. Inst. Maths. Applic. 26, no. 3). It is certainly not as easy to learn as MINITAB, nor is it so powerful as GLIM, but the combination of high-quality graphics and modern statistical procedures make it very

attractive for the research worker or as the basis of a Final Year or Postgraduate course in applied statistics. For lower level courses, that probably require only basics such as t- or F-tests, there is no good reason for changing to S-plus other than the very good graphics.

Bob Henery
bob@stams.strath.ac.uk

Products
Desktop Solutions
Enterprise Solutions
Specialized Add-Ons
Development Tools

Search
Go

INSIGHTFUL ANALYTIC SUITE

Home / **Products**

Desktop Solutions
Analytical software for PCs and UNIX workstations

S-PLUS for Windows: Statistical analysis, graphics and programming for Windows desktops

S-PLUS for UNIX: Statistical analysis, graphics and programming for UNIX and Linux workstations

Insightful Miner: Data mining for massive data sets using visual programming

Enterprise Solutions
Distributed analytics and information retrieval

S-PLUS Analytic Server: Enterprise distribution of analytics and Web-based decision support for Solaris and Linux servers

StatServer: Web-based decision support for NT and XP servers

Specialized Add-Ons
Optional components for Insightful products

S+NUOPT: Fast numerical optimization of functions of many variables

S+GARCH: Volatility modeling for financial time series data

S+SeqTrial: Design of group sequential clinical trials

EnvironmentalStats for S-PLUS: Statistical methods for analysis of environmental data

S-PLUS for ArcView GIS: Exploration and analysis of spatial data from within ESRI ArcView 3.2

S+Wavelets: Wavelet analysis of time series and image data

S+SpatialStats: Analysis of spatially correlated data

Related Information and Links

Products by Application Area

Success Stories

Insightful Solutions Library

Custom Solutions

Insightful Applications

Finance
Biotech and Pharmaceuticals
Manufacturing
Telecom
Business Intelligence
Data Mining
CRM Analytics
Research & Statistics
Academia

Development Tools
Products and features for creating analytic applications

CONNECT/C++: Embed S-PLUS analytics in C++ applications

CONNECT/Java: Integrate S-PLUS analytics with Java applications

S+SDK: Call S-PLUS functions from C applications

S-PLUS Graphlets: Interactive graphics on the Web

The S Language: The award-winning language for data analysis

Featured Products by Application Area

Statistical Analysis

S-PLUS for Windows
S-PLUS for UNIX
StatServer
S-PLUS Analytic Server

Data Mining

Insightful Miner

Business Analytics

S-PLUS Analytic Server
StatServer
S-PLUS Graphlets

Industry-Specific

S+SeqTrial
S+NUOPT
S+GARCH

Home | Products | Services | Partners/Alliances | Customers | Support | Company | Contact Us

Contact Us: info@insightful.com US Sales: 800 569 0123 UK Sales: +44 (0) 1276 450 111

Make Better Decisions Faster

S-PLUS®




- Statistics
- Data Mining
- Predictive Analysis
- Business Intelligence

Insightful™
intelligence from data

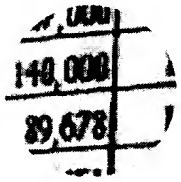
www.insightful.com

Discover Why S-PLUS is the Premier

For more than ten years, Insightful Corporation has been a leading supplier of data analysis solutions that drive better decisions faster by revealing patterns, trends, and relationships.

Critical business decisions depend on precise analysis of data. Industry-leading professionals from Fortune 1000 companies such as AT&T, Merrill Lynch, Merck Pharmaceutical and The Pillsbury Company rely on S-PLUS because it offers a flexible, interactive environment for analyzing, visualizing and presenting data.

S-PLUS, our flagship product, is one of the most powerful data analysis packages on the market. With S-PLUS you can streamline your data analysis process from accessing your data to sharing your results with colleagues or other decision-makers. From understanding your customers to ensuring product quality, S-PLUS gives you the tools to make better decisions today.



Award-winning S Language.
S-PLUS 6 is based on S version 4, an award-winning, object-oriented language developed at Lucent Technologies' Bell Labs specifically for data visualization, exploration and programming with data. The S System has been recognized with the prestigious Association for Computing Machinery Software Systems Award, (other recipients include UNIX, TCP/IP and Mosaic). With more than 4,200 statistical, graphical and programming functions built in, you can create applications in S in a fraction of the time it would take using lower-level languages like C or Visual Basic. The S language is platform-independent, so your applications will run on either Windows or Unix.

- **Data insight at your fingertips.**
  S-PLUS's intuitive graphical user interface offers the look-and-feel of Microsoft Office applications making it easy for you to access and analyze data. Embed graphs into Word or PowerPoint presentations with point-and-click ease and share your results with key decision makers.
- **Microsoft Excel integration improves data import and transfer capabilities.**
  You can open Excel worksheets within S-PLUS, perform analyses and create graphics directly from your data. Since your data stays in Excel, you won't spend time transferring results back and forth between Excel and S-PLUS.
- **Easily import and export data from Oracle, SAS, SPSS and other standard formats.**
  S-PLUS makes it easy to access data from virtually any source including Excel, SAS, SPSS and data bases including Oracle, Sybase and SQL Server. S-PLUS offers extensive import and export capabilities to help you move data from one file format to another.

Easily import and export data from the following formats

- SAS
- SPSS
- Excel
- Text (ASCII)
- Quattro Pro
- Paradox
- Lotus 1-2-3
- dBase
- Sigma Plot
- Systat
- STATA
- Gauss
- Access
- MATLAB
- EIM
- Bloomberg
- FAME
- Minitab
- FoxPro
- EpiInfo
- Informix
- Oracle
- Sybase
- SQL Server
- ODBC

Export graphics in the following fomats

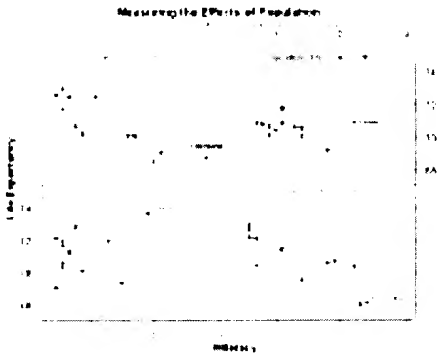
- Windows Bitmap (BMP)
- Encapsulated PostScript (EPS)
- CompuServe (GIF)
- GEM Bitmap (IMG)
- JPEG (JEG)
- Adobe Photoshop (PSD)
- Adobe PDF (PDF)
- HP Printer Control Language (PCL)
- PaintBrush (PCX)
- Tagged Image Format (TIF)
- True Vision Targa (TGA)
- Windows Metafile (WMF)
- Portable Network Graphics (PNG)

"The S System has forever altered the way people analyze, visualize and manipulate data"
Association for Computing Machinery (ACM)

www.insightful.com

Solution for Data Analysis

- **Share your discoveries using interactive graphics.**
  S-PLUS graphics are object-oriented so you can customize every level of detail to create the perfect graphic for your presentation. S-PLUS lets you interact with your data and identify unusual values or select relevant subsets. And S-PLUS's new Graphlets™ technology allows you to publish your graphics on the Web and give your readers the opportunity to interact with your data in real time.



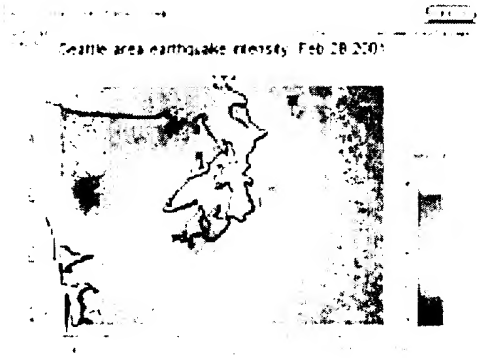
- **Visualize multi-dimensional data using Trellis graphics.**
  S-PLUS is the only data analysis package to offer Trellis graphics, a revolutionary way to visualize relationships in multi-dimensional data. Developed by researchers at Bell Labs, Trellis graphics help you discover hidden relationships in your data by computing graphical views sliced on one or more conditioning variables. No other graphing technique has as much power and flexibility.

- **Select from more publication-quality graphics and formats.**
  S-PLUS offers an extensive selection of 2D and 3D graph types. From histograms to bar charts to scatterplots, the graphics library in S-PLUS is truly comprehensive. Easily customize your graphs including line weights, colors and fonts for publication-quality graphics.

S-PLUS Graphlets™ offer real-time Web-based interactive graphics.

Introducing a new graphics format that allows you to add interactivity directly into graphics published on the Web. S-PLUS Graphlets offer the flexibility of the S-PLUS graphics engine to create exactly the graph you want, and then make it dynamic by allowing the viewer to drill-down into your data or create hyperlinks from your data points to other information or graphics located elsewhere on the Web. Examples of S-PLUS Graphlets can be found at www.insightful.com graphlets.

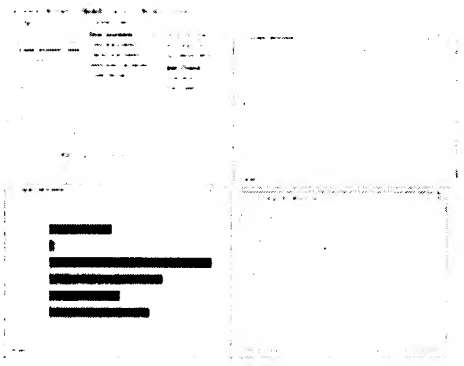
NEW! S-PLUS Graphlets
Now you can create interactive graphics where viewers can drill-down into the graphic to view information or linked Web pages.



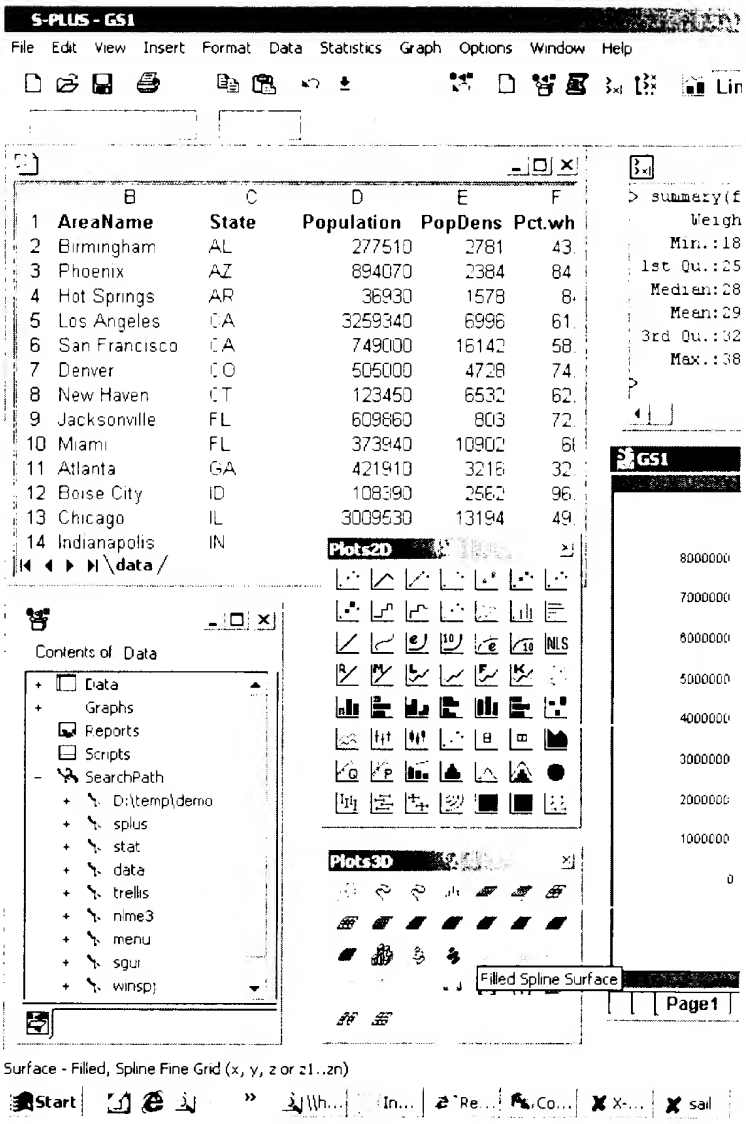
www.insightful.com

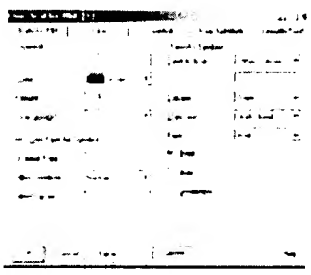
Visualize Your Data With S-PLUS

NEW! Excel Link Now you can open Excel worksheets from within S-PLUS. Perform analyses and create publication-quality graphics easily.



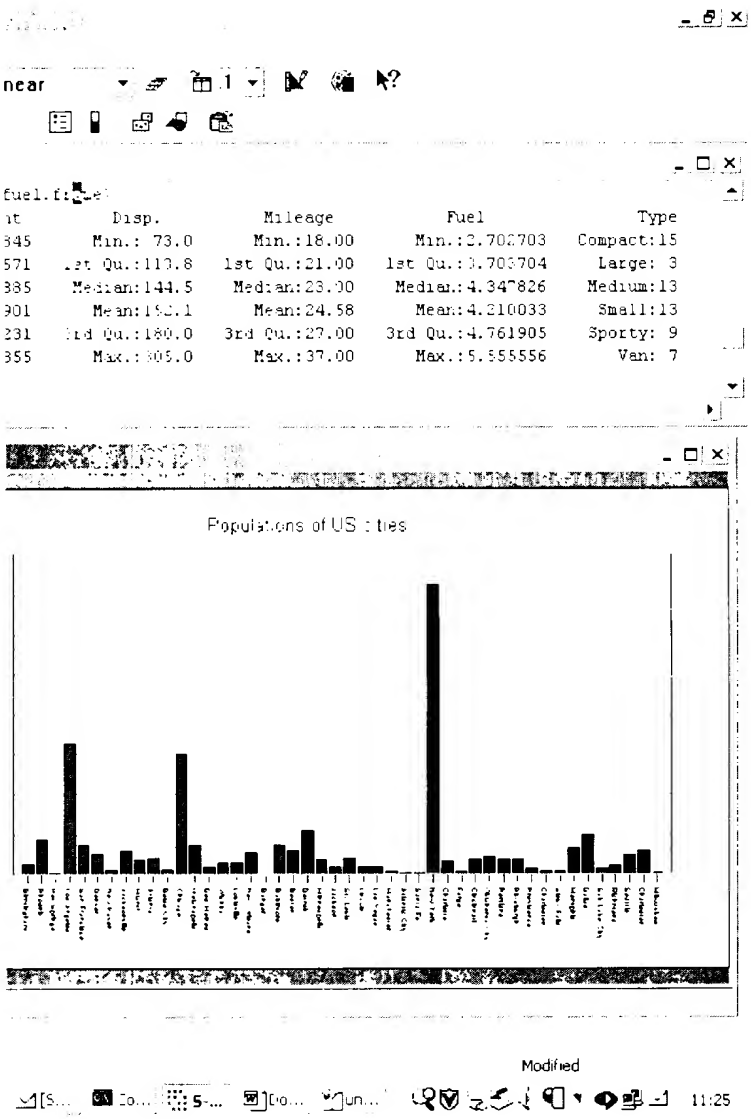
On Unix and Linux platforms, you can now access the powerful statistical and graphing techniques of S-PLUS through a graphical user interface. Easily import and export your data, run statistical analyses, and create revealing graphs, all through convenient menus and dialogs like those available on Windows. For programming, an interactive commands window gives you access to all the power and flexibility of the S language.





Change the Details of Your Graph With Ease Point-and-click control over every detail of your graphs makes it easy to produce stunning publication-quality output. Change line weights, axes, colors, labels, fonts, symbol types and more with ease.

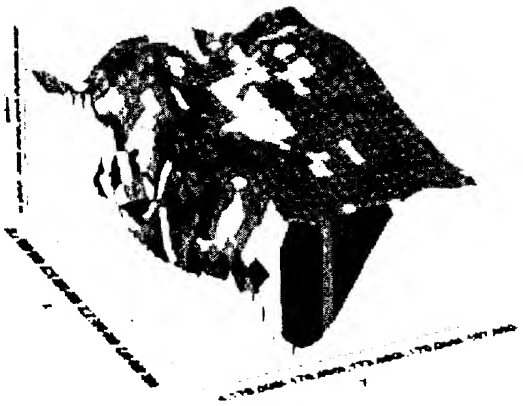
www.insightful.com

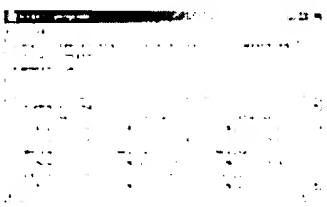


Completely Customize the User Interface
Change toolbars, menus and dialogs to suit your working style. Add and delete options with ease.

Produce Publication-Quality Output
Analyze, visualize and present your data using over 80 2D and 3D graph types.

Yosemite National Park Vegetation Map





Automate Repetitive Tasks With Powerful Scripting Capabilities
All S-PLUS operations are recorded in scripts which can be saved and executed to automate repetitive tasks. Drag-and-drop objects into a script window to instantly generate S-PLUS interface commands, and create your own buttons by dragging your scripts onto the toolbar. Share your work with your colleagues by giving them your toolbar and script files.

www.insightful.com

S-PLUS starts where your spreadsheet leaves off.

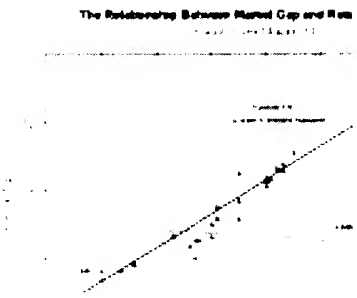
Spreadsheets are great for entering and organizing data, but if you're also trying to analyze your data with a spreadsheet you're potentially missing valuable insights by limiting yourself to inadequate methods. With S-PLUS 6 you get the best of both worlds—just open your Excel worksheet from within S-PLUS, select a region of data you wish to analyze, and instantly gain access to all of S-PLUS's advanced statistical and graphical functionality applied to your Excel data.

"...none has approached S-PLUS ... in terms of accuracy."
DM Review

- **Perform analysis using the most comprehensive solution available today.**
  S-PLUS offers over 4,200 built-in data analysis functions including modern and classical techniques. Convenient menus, toolbars and dialogs let you access and analyze your data easily.

- **Choose from the most comprehensive set of modern and robust methods available.**
  Tap into the power of linear and nonlinear regression, generalized linear models, generalized additive models, tree models, smoothing splines, survival analysis, time series, cluster analysis, robust methods, analysis of data with missing values, multiple comparisons and much, much more.

- **The S language gives you control over your data.**
  Precise, in-depth analysis may require extensive data manipulation and cleansing. With the S anguage at the core of S-PLUS you can easily prepare your data for analysis and graphing. From data transformations to data cleansing and validating data integrity, the S language provides the tools to get the job done efficiently, leaving you with a script file of transformations made to validate your actions.



- **Select the model that can provide you with the best results easily.**
  With the object-oriented S-PLUS environment, all functions, data and fitted models are handled as objects. This allows you to fit alternative models using both classical and modern methods so you can be confident the model you have selected will deliver the best results.

- **Create or extend analysis methods to meet your specific needs.**
  S-PLUS offers a powerful programming language allowing you to create or extend analyses. As your analyses become more complex, S-PLUS can be extended to meet the challenge. Tap into the power, flexibility and extensibility of S-PLUS to take your analyses to the next level.

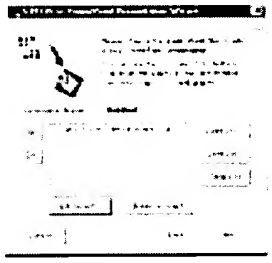
- **Cutting-edge methods available as Web downloads.**
  S-PLUS is the environment of choice for researchers world-wide developing advanced statistical methods for new problems in the world of data analysis. New S-PLUS functions and programs are available for download from third-party websites or from our own S-PLUS Community forum.

www.insightful.com

- **Create Powerpoint presentation slides with a click of a mouse.**
  S-PLUS graphics can be embedded in standard Microsoft Office products from PowerPoint slides to Word documents. Translate your results into professional looking reports using S-PLUS's unique graphic capabilities.

- **Total control over your results.**
  With S-PLUS you're not limited to pre-defined templates for your results. The S language gives you the freedom to present your results in any way you need: graphically, as free-form text, or as Web ready HTML tables.

- **Export your graphs into versatile formats for Web or print presentations.**
  Export publication-quality S-PLUS graphs to popular graphical formats including PostScript, GIF, PDF, and Windows MetaFile. Create interactive graphics on your web pages using Java-based S-PLUS Graphlets™.



S-PLUS: Power where you need it

S-PLUS has a completely open interface, allowing it to be integrated into virtually any system. Regular analyses can be automated using S-PLUS's batch processing system, employing the flexibility of the S command language. On Unix systems, S-PLUS's CONNECT/Java Interface allows S-PLUS to be integrated with any Java application. On Windows, the CONNECT/C++ Interface allows you to access S-PLUS's complete range of analytic methods from C++ applications you develop. And S-PLUS's DDE and OLE Automation interfaces allow you to integrate S-PLUS with other Windows applications, allowing you to access S-PLUS functionality within Excel or from Visual Basic applications.

Business Benefit

- **Leverage Statistical Expertise**
- **Deploy Analyses Quickly and Easily**
- **Access Analyses Through the Web**

"By launching a Web browser, employees can crunch data that analysts at headquarters once had to do"
Business Week

Distribute Analytics Enterprise-wide

Today's decision-makers need access to real-time analysis of business data to anticipate risk, predict customer behavior and capitalize on emerging market opportunities. Insightful's client-server products can provide decision makers with up-to-the-minute results on their desktop, based on analytic or graphical methods using S-PLUS. With a custom-created application designed expressly for the problem at hand, decision makers gain insight from their data soon- [illegible] them to make informed decisions and pursue [illegible] with a strategic advantage. [illegible] www.insightful.com/integration

S-PLUS Comprehensive Feature List

STATISTICAL & NUMERICAL TECHNIQUES

Basic Statistics

- [illegible]

Hypothesis Tests and Confidence Intervals

- [illegible]

Regression

- Basic linear regression
- Polynomial regression
- Model diagnostics
- Prediction and confidence intervals
- Stepwise selection of models
- Parametric spline models
- Least trimmed squares regression
- Constrained regression
- Logistic regression
- Generalized linear models
- Maximum likelihood robust regression
- Robust MM regression

Analysis of Variance

- Flexible specification of variables (interactions, nesting, transformations)
- Automatic generation of dummy variables
- Choice of contrasts
- Type III sums of squares
- Rank tests (Kruskal-Wallis, Friedman)
- Designed experiments (one-way, two-way, tests on split plot imbalances)
- Variance component estimation
- Multivariate ANOVA
- Multiple comparisons (Fisher, Tukey, Dunnett, Sidak, Bonferroni, Scheffe, simulation based)

Nonlinear Regression and Maximum Likelihood

- Nonlinear regression
- Nonlinear mixed-effects models
- Quasi-likelihood
- Constrained nonlinear regression

Nonparametric Regression

- [illegible]

Tree models

- [illegible]

Smoothing

- [illegible]

Linear and Nonlinear Mixed-Effects Models

- [illegible]
- Repeated measures models
- Automated algorithms for hierarchical models
- Flexible covariance matrix structures
- Built-in functions for linear mixed effects models for Gaussian responses and first order compartment models for parameter logistic
- Virtually any user-defined model can be easily handled

Resampling

- Bootstrap
- Jackknife

Multivariate Analysis

- Canonical correlation
- Discriminant analysis
- Factor analysis
- Multidimensional scaling
- Principal components
- MANOVA

Cluster Analysis

- K-means
- Partitioning around medoids
- Monothetic clustering
- Model based clustering
- Crisp and fuzzy clustering
- Divisive and agglomerative methods

Quality Control

- Shewhart chart
- Cusum chart
- Charts based on quantile [illegible]

Power and Sample Size

- Normal mean
- Binomial proportion

Survival Analysis

- Kaplan Meier curves
- Cox proportional hazards models
- Left, right, and interval censoring
- Time dependent covariates and strata
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]

Time Series Analysis

- [illegible]
- [illegible]
- Box-Jenkins ARIMA models
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]

Date, Time, and Calendar Data

- [illegible]
- Subsetting and operations
- Aligning, merging, merging information
- Times and dates from many sources, conventions
- Time zones with daylight savings time and world
- Holidays and market calendars
- Flexible time and date format
- Business quarterly dynamics with special holiday sets that span years
- Relative time, time sequence and event objects

Mathematical Computations

- Vector and matrix computations
- Matrix decompositions
- Systems of linear equations
- Eigenvalues
- Nonlinear optimization
- Constrained optimization
- Ordinary differential equations
- Numerical integration

Robust Methods

- New S language for developing robust analysis functions, methods
- New plots for outlier detection and comparisons
- New robust models and comparison paradigm

Missing Data Library

- Multiple imputation
- Gaussian, logistic, and conditional Gaussian models

Large Data Set Support

- Memory mapping
- Reference counting
- Command language interface call to C
- Variable selection on import (keep, drop)
- Support for sequential processing

GRAPHICS & VISUALIZATION

Plot Type Highlights

- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]

Advanced Data Visualization

- Exclusive new sequence of colors, symbols, and line attributes for enhanced visualization
- Trellis display for optimal aspect ratios
- Multiple simultaneous 3D rotation views
- 2D projection in 3D space
- Multiple graphs per page with automatic scaling
- Brush and slice dynamic visualization
- Interactive observation identification of points
- Multiple user-defined color maps
- Graphpads

Customizability and Editing

- Flexible page layout
- Overlay graphs or display side by side
- Variety of fonts and symbols
- Control over line style, marker type, colors, labels, tick marks, text, fill pattern
- Multiple independently scaled x and y axes
- Multiple line text annotation

Import and Export

- Import SAS, SPSS, Excel, Matlab, and other file formats
- Query Oracle, Sybase, Informix, and ODBC databases
- Export graphics as PDF, postscript or WPG

PROGRAMMING & EXTENSIBILITY

Object-Oriented Language

- Uses the object oriented S programming language
- Over 4200 built-in functions
- Users may modify functions and write new functions
- Rich data structures include vectors, matrices, arrays, and lists

- [illegible]
- [illegible]
- [illegible]
- [illegible]

Interconnectivity

- [illegible]
- [illegible]
- [illegible]
- Standard GUI classes
- Access operating system utilities
- [illegible]
- [illegible]

MS Office Integration (Windows)

- [illegible]
- Embed S-PLUS graphics in Word and edit them in place
- Create PowerPoint slides from S-PLUS graphics automatically
- User-defined SPSS wizards to create S-PLUS graphs from within Excel or SPSS

User Contributed Code

- Libraries associated with the book *Modern Applied Statistics with S-PLUS* by Venables and Ripley
- Biostatistical and epidemiological modeling by Harrell

Help and Documentation

- Extensive documentation
- Comprehensive Index and Search
- Technical and reference guidelines
- Self documenting objects

60-DAY MONEY-BACK GUARANTEE

Call now to discuss your business needs

Insightful
intelligence from data

1700 Westlake Avenue North Suite 500
Seattle WA 98109
Tel: 206.283.8802 • 800.569.0123
Fax: 206.283.6310
email: info@insightful.com
www.insightful.com

International Division

Knightway, House Park Street
Bagshot, Surrey GU19 5AQ, UK
Tel: +44 1276 45[illegible]
Fax: +44 1276 451224
email: info-uk@insightful.com
www.insightful.com

Editors Choice